Received: 26 April 2017 | Revised: 27 December 2017 | Accepted: 14 January 2018

DOI: 10.1002/sim.7625

RESEARCH ARTICLE

WILEY Statistics in Medicine

Addressing the estimation of standard errors in fixed effects meta-analysis

Clara Domínguez Islas[1,2] | Kenneth M. Rice[3]

[1]Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[2]MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK

[3]Department of Biostatistics, University of Washington, Seattle, WA, USA

Correspondence
Clara Domínguez Islas, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., M2-C200, PO Box 19024, Seattle, WA 98109-1024, USA.
Email: cdomingu@fredhutch.org

Funding information
Department of Biostatistics, University of Washington and the MRC Biostatistics Unit, Grant/Award Number: U105260558

Standard methods for fixed effects meta-analysis assume that standard errors for study-specific estimates are known, not estimated. While the impact of this simplifying assumption has been shown in a few special cases, its general impact is not well understood, nor are general-purpose tools available for inference under more realistic assumptions. In this paper, we aim to elucidate the impact of using estimated standard errors in fixed effects meta-analysis, showing why it does not go away in large samples and quantifying how badly miscalibrated standard inference will be if it is ignored. We also show the important role of a particular measure of heterogeneity in this miscalibration. These developments lead to confidence intervals for fixed effects meta-analysis with improved performance for both location and scale parameters.

KEYWORDS

fixed effects, heterogeneity, meta-analysis, random effects

1 | INTRODUCTION

Meta-analysis, "the use of statistical methods to summarize the results of independent studies,"[1,2] is a pivotal component of systematic reviews[2] that have been extensively used to synthesize the increasing amount of evidence produced in health care research.[3] In broad terms, the primary aim of most meta-analyses is to make some form of inference on the size of effects across several similar studies. A typical goal is to summarize all the studies and make inference on the magnitude and direction of some form of average effect. Measures of spread, ie, how the study effects vary across different studies, are also often considered.

A recent review paper[4] notes that the well-known inverse-variance fixed effects estimate, which can be easily motivated as an estimate of a "common effect," can also be interpreted as a particular average of study-specific effects, without any requirement that study effects be homogeneous. However, using this alternative interpretation is only straightforward when the standard errors are known with negligible error, a simplifying assumption that is rarely entirely plausible in practice. The impact of this assumption has been studied in some special case (ie, when assuming homogeneity),[5,6] but a general understanding of how fixed effects meta-analysis is affected is missing from the literature. In the present work, we therefore provide several tools to do statistical inference for fixed effects meta-analysis when this assumption cannot be made.

The paper is structured as follows: in Section 2, we review the different statistical models that can be used for meta-analysis, their parameters of interest, and popular estimation methods. In Section 3, we show how the precision weighted average effect arises naturally when considering optimal summary measures and also propose a measure of het-

erogeneity around this parameter. In Section 4, we present intuitive and formal arguments for why the impact of standard error estimation does not go away with larger sample sizes and how this impact depends on the underlying heterogeneity. We present simulation results comparing several confidence intervals for the precision weighted average that allow for estimation of standard errors, and for a related measure of heterogeneity. Finally, in Section 5, we give an applied example to illustrate and compare the different approaches to meta-analysis and conclude with a discussion in Section 6.

2 | REVIEW OF APPROACHES TO META-ANALYSIS

In this section, we describe 3 different approaches to meta-analysis, in which different assumptions are made about the underlying true effect size parameters in the studies. Table 1 provides a summary of these approaches, and we subsequently present further details on the precision weighted average (Section 2.1), testing and quantifying heterogeneity (Section 2.2), and random effects analysis (Section 2.3).

The first approach is the *fixed effect* (singular) meta-analysis, also called the *common effect* meta-analysis.[4] This approach is based on the assumption of a single, common effect underlying all studies.[6] Under this simplifying assumption that all study effects are identical, the average effect is equivalent to the common effect size estimated in each study. Although commonly used, this method has often been judged inadequate in practice, as effects from different studies are expected to differ given the variability in study design, population, interventions, etc.[7-9]

A second approach is the *fixed effects* (plural) meta-analysis, based on the assumption that the effects underlying the studies at hand are unknown, but fixed, and not necessarily identical.[10,11] Using the fixed effects approach, it is common to estimate the inverse-variance weighted average of the studies' effect sizes,[4] but estimation of other weighted averages is also possible.[11,12] As recently discussed by Rice et al,[4] the inverse-variance weighted average estimates a reasonable and interpretable parameter, even when the effect sizes are assumed to be different, but it may be a somewhat incomplete summary of the effect sizes if they are too heterogeneous.[13]

The third approach is the *random effects* meta-analysis, where the effect-size parameters are considered to be a random sample from a population, ie, they follow a probability distribution.[14] By using random effects as a sampling model, this analysis allows the estimation of the average effect size in the population of effect sizes one might ever have observed.[14] (Details are given in Section 2.3.) This method not only takes into account the heterogeneity between studies but also provides a natural way of quantifying it,[15] making it a more attractive choice over the common and fixed effects approaches.[7,8] On the other hand, as pointed out by Higgins et al,[15] this approach is based on a construct of an hypothetical population of studies or study effects, so the interpretation of the analysis is potentially unclear and confusing. The relevance of random effects analyses that focus on the mean of a population of study effects has been questioned.[16] An alternative derivation for the random effects approach motivates the distribution of effect sizes not as a sampling distribution, but arising from a priori exchangeability in a Bayesian analysis—or an approximately Bayesian analysis, as noted in Higgins et al.[15]

In each of the 3 approaches described, appeals to some form of frequentist optimality can be made. In the common effect approach, when the study-specific standard errors are known precisely, the optimality is straightforward; without

TABLE 1 Statistical assumptions from 3 different approaches to meta-analysis of k studies, their target parameters for location summary and estimators

Common assumption	$\hat{\beta}_i \sim N(\beta_i, \sigma_i^2)$, with σ_i^2 known, for $i = 1, 2, \ldots, k$		
Approach-specific	**Common Effect**	**Fixed Effects**	**Random Effects**
assumption	$\boldsymbol{\beta}_i = \beta_0 \forall i, \beta_0 \in \mathbb{R}^k$	$\boldsymbol{\beta} = (\beta_i, \ldots, \beta_k)^T \in \mathbb{R}^k$	$\beta_1, \ldots, \beta_k$ iid $f(\mu, \tau^2)$
Inference target	β_0	$\beta_F = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}\beta_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}$	μ
Estimator	$\hat{\beta}_0 = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}\hat{\beta}_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}$	$\hat{\beta}_F = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}\hat{\beta}_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}$	$\hat{\mu} = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2+\hat{\tau}^2}\hat{\beta}_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2+\hat{\tau}^2}}$
Standard error	$\widehat{\mathrm{SE}}(\hat{\beta}_0) = \sqrt{\frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}}$	$\widehat{\mathrm{SE}}(\hat{\beta}_F) = \sqrt{\frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i^2}}}$	$\widehat{\mathrm{SE}}(\hat{\mu}) = \sqrt{\frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i^2+\hat{\tau}^2}}}$
Heterogeneity	Not present, by assumption	Hypothesis test based on $Q = \sum_{i=1}^k \frac{1}{\sigma^2}(\hat{\beta}_i - \hat{\beta}_F)^2$	$\hat{\tau}^2 = \max\left\{0, \frac{Q-(k-1)}{\sum \sigma_i^{-2} - \sum \sigma_i^{-4}/\sum \sigma_i^{-2}}\right\}$
Consistency	Not evaluated	$I^2 = \frac{Q-(k-1)}{Q}$	$I^2 = \frac{Q-(k-1)}{Q}$

any distributional assumptions, the inverse-variance weighted estimator provides the best linear unbiased estimator of the common effect, or the unique minimum variance unbiased estimator under the further assumption of normality of effects estimates.[6] When the study-specific standard errors must be estimated, a normal approximation based on the asymptotic distribution of the estimator is commonly used[17, chapter 6]; in the common situation where all the studies are large, the standard errors are known with great accuracy, and any nonasymptotic inefficiency is extremely minor.

For the fixed effects approach, it has been shown by Lin and Zeng[18] that the analysis provides, in many situations, a statistically efficient estimate of the parameter that would be estimated, were it possible to pool the data across studies and to perform a single regression analysis that adjusts for study. However, this pooling is inherently somewhat hypothetical; were it possible to do it, there would often be little motivation for use of meta-analysis, and so it may not always be obvious that this parameter is of direct interest.

The random effects approach has perhaps the least direct connection to optimality, while likelihood-based and fully Bayesian methods in general have guarantees of good large-sample properties, under correct model assumptions,[19,20] in finite samples or when the model is misspecified there are no such guarantees. Indeed, the finite-sample sensitivity of Bayesian random effects meta-analysis to choice of priors is well documented[21-24] and is a cause for concern in practice.[25], chapter 5

2.1 | The precision weighted average

Let $\beta_1, \beta_2, \ldots \beta_k$ be the true effect sizes from k different studies and let $\hat{\beta}_i$ be the estimate of the true effect β_i, with corresponding standard error σ_i, which we assume known for now. The *precision weighted average* or *inverse-variance weighted average* of the true effect sizes is

$$\beta_F = \frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\beta_i}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}. \tag{1}$$

This parameter is a quantity of interest in either common effect or fixed effects meta-analysis; under the common effect model, β_F reduces to the common effect β_0 seen in Table 1; under the fixed effects model, β_F is a weighted average of the effect-sizes β_i, where the weight is proportional to the precision with which each effect size can be estimated, giving more weight to those that can be estimated more precisely

If the study-specific standard errors σ_i are assumed to be known, a natural estimator of β_F is given by

$$\hat{\beta}_F = \frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\hat{\beta}_i}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}, \text{with} \operatorname{SE}(\hat{\beta}_F) = \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}}. \tag{2}$$

Optimality of β_F under a common effect approach has already been mentioned. In fixed effects meta-analysis with known standard errors σ_i, $\hat{\beta}_F$ directly inherits any efficiency properties from the $\hat{\beta}_i$'s, as would any linear combination of the effect-size estimates. This means that $\hat{\beta}_F$ is an unbiased, efficient, and/or normally distributed estimator of β_F, if within each study, the estimator $\hat{\beta}_i$ can be assumed to be an unbiased, efficient, and/or normally distributed estimator of β_i.

Confidence intervals for β_F are usually derived from a normal approximation, appealing to the large sample properties of the study estimators. Transformations of the outcome measure have also been recommended, such as normalizations, log transformations, bias corrections,[17] and/or variance stabilizing transformations.[26,27] The small sample properties of the normal approximation and sensitivity to the assumption of known variances have been studied through simulation studies,[17,28,29] and some corrections and tests based on more robust test statistics have been proposed.[5,29]

2.2 | Testing homogeneity and quantifying heterogeneity

In both fixed effects and common effect work, it is common to test *homogeneity* of the study effects, that is, to test the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$, against the general alternative of *heterogeneity*, where some β_i are not equal. In common-effect meta-analysis, this test assesses a key modeling assumption, while in the fixed effects analysis, the test simply gives a statistical measure of how much heterogeneity is present.

When assessing homogeneity, a commonly used test statistic is

$$Q = \sum_{i=1}^{k} \frac{1}{\sigma_i^2}(\hat{\beta}_i - \hat{\beta}_F)^2.$$

Under normality of the effect estimates ($\hat{\beta}_i \sim N(\beta_i, \sigma_i^2)$), Q is distributed noncentral chi-squared with $k-1$ degrees of freedom and noncentrality parameter

$$\lambda = \sum_{i=1}^{k} \frac{1}{\sigma_i^2}(\beta_i - \beta_F)^2,$$

with β_F as in (1). Q is independent of the $\hat{\beta}_F$ statistic,[4] which may simplify its interpretation.

Under the null hypothesis that all the effects are identical, the Q statistic is distributed central chi-squared with $k-1$ degrees of freedom, thus providing a reference distribution to perform a test of homogeneity of effects. However, it also has been found that this test of homogeneity has low power when there are few studies[30] and is not adequate to summarize the extent of the heterogeneity present.[31]

Other statistics have been proposed to not only test but also evaluate the impact of the observed heterogeneity, and thus provide a better measure of the consistency between trials.[31] Although these measures have been motivated and derived from a random effects framework, they still have valid interpretation under a fixed effects framework.[4] The most-frequently used of these quantities is I^2, which can be calculated as

$$I^2 = \frac{Q-(k-1)}{Q},$$

and is interpreted as "the percentage of total variation across studies that is due to heterogeneity rather than chance."[32]

2.3 | Inference in random effects meta-analysis

Random effects meta-analysis is based on the assumption that the true study effects $\beta_1, \beta_2, \ldots \beta_k$ are an independent and identically distributed sample from some distribution. The inference is then focused on the parameters of this distribution, typically its mean (μ) and variance (τ^2).

With no further assumptions on the distribution of the random effects, an inverse-variance weighted average estimate of μ can be obtained,[14,15] along with an estimate of its standard error:

$$\hat{\mu} = \frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^2+\hat{\tau}^2}\hat{\beta}_i}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2+\hat{\tau}^2}}, \text{ with } \widehat{\text{SE}}(\hat{\mu}) = \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2+\hat{\tau}^2}}}. \tag{3}$$

The weights here involve both the within-study variance σ_i^2 and the heterogeneity (or between studies) variance τ^2, for which a moment-based estimator is

$$\hat{\tau}^2 = \max\left\{0, \frac{Q-(k-1)}{\sum \sigma_i^{-2} - \sum \sigma_i^{-4} / \sum \sigma_i^{-2}}\right\}.$$

As given here, $\hat{\mu}$ and $\hat{\tau}$ are known as the DerSimonian-Laird estimator for random effects meta-analysis.[14] Other similar moment-based estimator have been proposed.[33,34]

Under the further assumption that the study effects follow a normal distribution, maximum likelihood[35,36] and restricted maximum likelihood[37,38] methods can be used to obtain estimates of τ^2 and μ. Although these methods are iterative and do not provide closed form estimates, it should be noticed that both the maximum likelihood and REML estimators of μ take the same form as in (3). A simpler, noniterative method for estimating τ^2 has recently been proposed[39] and is also based on the assumption of a normal distribution of the study effects. The performance of the different estimation methods has been evaluated and compared, in terms of bias and efficiency,[34] as well as coverage probability.[40]

3 | UNDERSTANDING HOW ESTIMATION OF STANDARD ERRORS AFFECTS FIXED EFFECTS META-ANALYSIS

In Equation 1 of Section 2.1, we saw how the underlying parameter estimated by fixed effects meta-analysis is typically defined, in terms of standard errors. When the standard errors are not known but only estimated, this leaves the target of this analysis without a full definition. In Section 3.1, we provide a more concrete motivation, showing how the parameter estimated is optimal for inference, in a certain sense. The impact of estimated standard errors on this inference is explored in Section 3.2, and we see how this motivates the study of a particular scale parameter, describing heterogeneity, in Section 3.3.

3.1 | A location parameter for optimal estimation

Ideally, the parameters to which inference is targeted should be determined entirely by scientific criteria, ie, by research goals. But in practice these goals may not be known precisely enough to determine a single parameter for inference. In this situation, it makes sense to use statistical criteria to choose from among parameters that meet general research goals. In meta-analysis, where the general goal is to summarize study effects β_i by some form of average, we choose to pick the the affine combination (ie, the weighted average) of the β_i that can be most precisely estimated. This can also be stated as selecting the parameter for which the data provides the most information.

The main result here follows from a more general lemma, proved in Appendix A:

Lemma 1. *Let* $\{\boldsymbol{v}^T\boldsymbol{\beta} : \boldsymbol{v} \in \mathbb{R}^k, \boldsymbol{v}^T\mathbf{1}_k = 1\}$ *be the set of all possible affine combinations of the vector of effect-size parameters* $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)^T$ *and let* $\hat{\boldsymbol{\beta}}$ *be the vector of estimates* $(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)^T$ *with covariance matrix* $\boldsymbol{\Sigma}$*. Then the affine combination of the parameter vector* $(\boldsymbol{w}^T\boldsymbol{\beta})$ *for which the corresponding estimator* $(\boldsymbol{w}^T\hat{\boldsymbol{\beta}})$ *has the minimum variance is given by*

$$\boldsymbol{w} = \underset{\boldsymbol{v}:\boldsymbol{v}^T\mathbf{1}_k=1}{argmin} \left[\boldsymbol{v}^T\Sigma\boldsymbol{v}\right] = \frac{\Sigma^{-1}\mathbf{1}_k}{\mathbf{1}_k^T\Sigma^{-1}\mathbf{1}_k}$$

with $\mathrm{Var}(\boldsymbol{w}^T\hat{\boldsymbol{\beta}}) = (\mathbf{1}_k^T\Sigma^{-1}\mathbf{1}_k)^{-1}$.

In fixed effects meta-analysis, where the studies are independent, the covariance matrix of $\hat{\boldsymbol{\beta}}$ reduces to a diagonal matrix: $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_i^2\}$. From Lemma 1 and assuming that σ_i^2 is known exactly from each study, then the best affine combination of the effect-size parameters is

$$\left(\frac{\mathbf{1}_k^T\mathrm{diag}\{\sigma_i^{-2}\}}{\mathbf{1}_k^T\mathrm{diag}\{\sigma_i^{-2}\}\mathbf{1}_k}\right)\boldsymbol{\beta} = \frac{\sum_{i=1}^k \frac{1}{\sigma_i^2}\beta_i}{\sum_{i=1}^k \frac{1}{\sigma_i^2}} = \beta_F, \tag{4}$$

the precision weighted average of the effect-size parameters.

In the situation where the σ_i^2 are assumed known, the corresponding estimate $\hat{\beta}_F$ can be easily constructed, and used as described in Section 2.1.

But the same optimality of β_F holds even when the σ_i are not known. To show this formally, we express σ_i^2 as

$$\sigma_i^2 = (n_i\phi_i)^{-1} = N^{-1}(\eta_i\phi_i)^{-1},$$

where n_i and ϕ_i are the sample size and the Fisher information from each subject on β_i, respectively, in the ith study, N is the total sample size across all studies, and $\eta_i = n_i/N$ is the proportion of the total sample drawn from study i. Then formally, under the asymptotic regime where η_i are fixed when we consider larger N (ie, the same assumptions as in the earlier work of Lin and Zeng,[18] and indeed most asymptotic work), then the limiting value of the covariance matrix is $\boldsymbol{\Sigma} = N^{-1}\mathrm{diag}\{(\eta_i\phi_i)^{-1}\}$, and canceling terms in N, we find

$$\beta_F = \frac{\sum_{i=1}^k n_i\phi_i\beta_i}{\sum_{i=1}^k n_i\phi_i} = \frac{\sum_{i=1}^k \eta_i\phi_i\beta_i}{\sum_{i=1}^k \eta_i\phi_i}.$$

This shows that, without further assumptions, in large samples, β_F is the weighted average of the β_i parameters that can be most precisely estimated. When the true standard errors σ_i are not known but instead estimated by s_i, β_F can be

consistently estimated by a "plug-in" version of $\hat{\beta}_F$ from Equation 2. We denote this estimate as

$$\hat{\hat{\beta}}_F = \frac{\sum_i^k \frac{1}{s_i^2}\hat{\beta}_i}{\sum_i^k \frac{1}{s_i^2}}.$$

3.2 | Impact of estimated standard errors on $\hat{\hat{\beta}}_F$ with estimated standard errors

When using the precision weighted average from Equation 2, it is common to assume that the sample size in each study is large enough for the variance of the effect estimate (σ_i^2) to be approximated with negligible error by its estimate (s_i^2),[41] basing tests statistics and confidence intervals on the following plug-in estimator:

$$\hat{\hat{\beta}}_F = \frac{\sum_i^k \frac{1}{s_i^2}\hat{\beta}_i}{\sum_i^k \frac{1}{s_i^2}}, \text{ with } \widehat{\text{SE}}(\hat{\hat{\beta}}_F) = \sqrt{\frac{1}{\sum_{i=1}^k \frac{1}{s_i^2}}}. \tag{5}$$

The properties of Equation 5 in small sample size settings have been studied via simulation, with inflated type I error rates observed for the test of the null hypothesis $H_0 : \beta_F = 0$, due to underestimation of the standard error of $\hat{\hat{\beta}}_F$.[17,28,29]

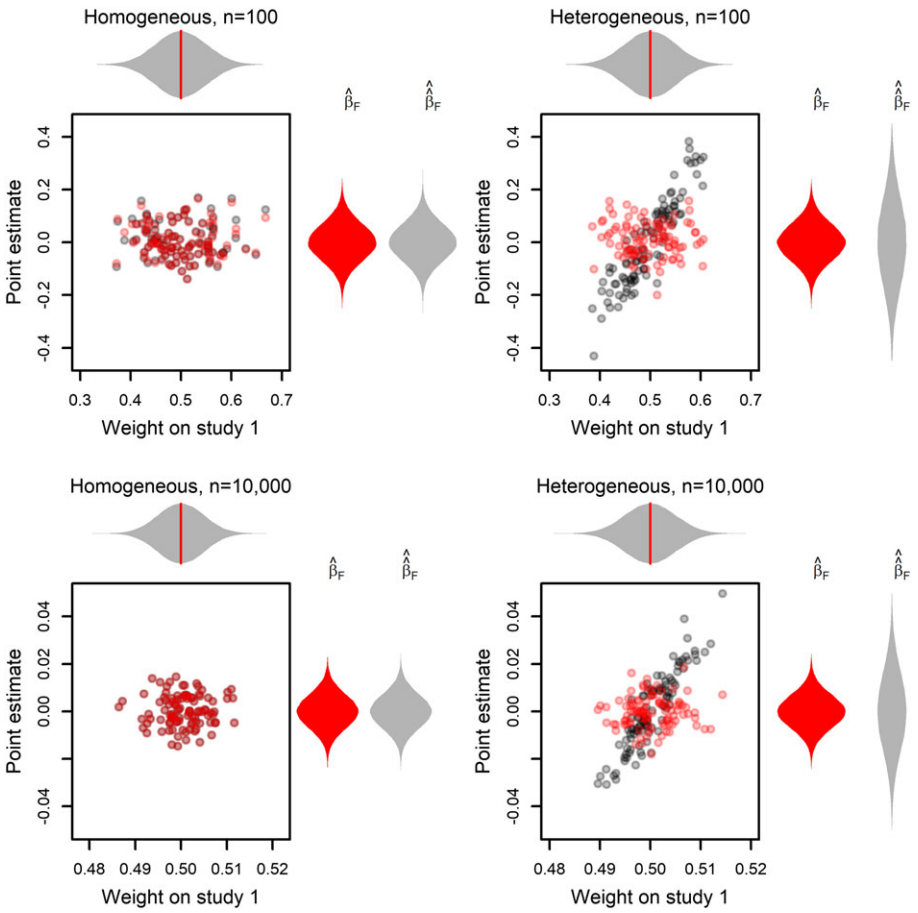

FIGURE 1 Comparison of the distributions of $\hat{\beta}_F$ and $\hat{\hat{\beta}}_F$ in a simple meta-analyses of 2 homogeneous studies with effect sizes $\beta_1 = \beta_2 = 0$ (*left column*) and 2 heterogeneous studies with effect sizes $\beta_1 = 1.5$ and $\beta_2 = -1.5$ (*right column*). We consider medium size studies with N=100 (*top row*) and very large studies with N=10 000 (*bottom row*). The y-axis and the vertical violin plots show the distributions of the estimates $\hat{\beta}_F$ with no uncertainty in the study weights (*in red*) and $\hat{\hat{\beta}}_F$ with estimated study weights (*in gray*). The x-axis and the horizontal violin plot show the distribution of the estimated weight given to study 1 in $\hat{\hat{\beta}}_F$. Notice that this same x-coordinate is used for both the gray and red diamonds, to illustrate their variability and their overlap (in the homogeneous case), but the weight for study 1 in $\hat{\beta}_F$ is always exactly 0.5, as indicated by the red line in the horizontal violin plot

Corrected and alternative test statistics have been proposed,[5,6,29] but all of them are based on the assumption of a common effect.

However, in our experience, many investigators expect that the effect of plugging-in s_i for σ_i should, in large samples, be negligible for inference on β_F, regardless of the underlying β_i—and so the simulation results can be ignored when studies have large sample sizes. This intuition appears to be based on experience with other small-sample corrections that change standard error estimates by factors of $n/(n-1)$ or $n/(n-p)$, which can be ignored with large n. However, this intuition does not apply to $\hat{\hat{\beta}}_F$; not only does the effect of plugging-in s_i remain at any sample size, its impact depends importantly on the heterogeneity between the various β_i.

To better understand how the potential heterogeneity affects the estimation of $\mathrm{Var}[\hat{\hat{\beta}}_F]$, we decompose the variance of $\hat{\hat{\beta}}_F$ as

$$\begin{aligned}\mathrm{Var}[\hat{\hat{\beta}}_F] &= \mathrm{E}[\mathrm{Var}(\hat{\hat{\beta}}_F|s_1^2, \ldots, s_k^2)] + \mathrm{Var}[\mathrm{E}(\hat{\hat{\beta}}_F|s_1^2, \ldots, s_k^2)] \\ &= \mathrm{E}\left[\left(\sum_i^k \frac{\sigma_i^2}{s_i^4}\right)\Big/\left(\sum_i^k \frac{1}{s_i^2}\right)^2\right] + \mathrm{Var}\left[\left(\sum_i^k \frac{1}{s_i^2}\beta_i\right)\Big/\left(\sum_i^k \frac{1}{s_i^2}\right)\right],\end{aligned} \tag{6}$$

where the second line follows from assumptions that each $\hat{\beta}_i$ is unbiased, is independent of its corresponding standard error estimate s_i, and has variance σ_i^2.

Under exact homogeneity, the second term in Equation 6 simplifies to zero, but is otherwise strictly positive. Moreover, this second term does not become small compared with the first term at larger sample sizes. Before showing this phenomenon formally, we first illustrate it in Figure 1. It shows a simple fixed effects meta-analysis of just 2 studies, of equal sample size, precision, but potentially with unequal β_i. Comparing behavior of the fixed effects estimate with known standard errors ($\hat{\beta}_F$, in red) and estimated standard errors ($\hat{\hat{\beta}}_F$, in gray), we see that for heterogeneous data, regardless of sample size, the estimated standard errors give a more variable estimate. This is because $\hat{\hat{\beta}}_F$ is "tilted" closer to β_1 or β_2 when—by chance alone—study 1 or 2 receives greater weight. This pattern persists at larger sample sizes, so while the absolute amount of extra noise induced is reduced, the relative variabilities remain essentially unchanged. For the homogeneous settings, the 2 β_i are equal, so no "tilting" occurs, but for the heterogenous settings, the precisions differ by a factor of more than 5.

To build further intuition about the extra variability induced by using estimated standard errors, we now provide an analytic version of the results illustrated in Figure 1. To do this, we write the variance of each $\hat{\beta}_i$ as $\mathrm{Var}(\hat{\beta}_i) = \sigma_i^2 = (n_i\phi_i)^{-1}$, and the estimator s_i^2 of σ_i^2 as $s_i^2 = (n_i\hat{\phi}_i)^{-1}$, so that

$$\hat{\hat{\beta}}_F = \frac{\sum_i^k \frac{1}{s_i^2}\hat{\beta}_i}{\sum_i^k \frac{1}{s_i^2}} = \frac{\sum_i^k n_i\hat{\phi}_i\hat{\beta}_i}{\sum_i^k n_i\hat{\phi}_i}. \tag{7}$$

Additionally, we make the large-sample approximation that each $\hat{\phi}_i$ is asymptotic normal, with asymptotic variance given by some function of the distributional moments of the population(s) in study i, that we write as $f_i(\boldsymbol{\theta}_i)$.*. Using the usual assumptions of normality of $\hat{\beta}_i$ and independence of $\hat{\beta}_i, s_i$, then by the delta method, we obtain

$$\sqrt{N}\left(\hat{\hat{\beta}}_F - \beta_F\right) \to_d N\left(0, \frac{1}{\left(\sum_i^k \eta_i\phi_i\right)}\left[1 + \frac{\sum_i^k \eta_i(\beta_i - \beta_F)^2 f_i(\boldsymbol{\theta}_i)}{\sum_i^k \eta_i\phi_i}\right]\right). \tag{8}$$

Details are provided in Appendix B. Comparing Equation 8 to the standard error in Equation 5, we see that the asymptotic variance of $\hat{\hat{\beta}}_F$ is the product of the asymptotic variance when the variances are known multiplied by an *inflation factor*, given in square brackets. This inflation factor, which accounts for the uncertainty in the estimation of the standard errors, depends on the squared deviations of β_i from β_F, and thus, it will reduce to 1 under homogeneity but will increase as the dispersion of the effect sizes increases. We also notice that the squared deviations are multiplied by $f_i(\boldsymbol{\theta}_i)$,

*The specific form of $f_i(\boldsymbol{\theta}_i)$ will depend on the type of estimator used, the study's randomization ratio as well as the variances and kurtoses of the treatment and control subpopulations. For this reason, we have decided to use this generic expression but have also provided detailed case-specific derivations in Appendix E

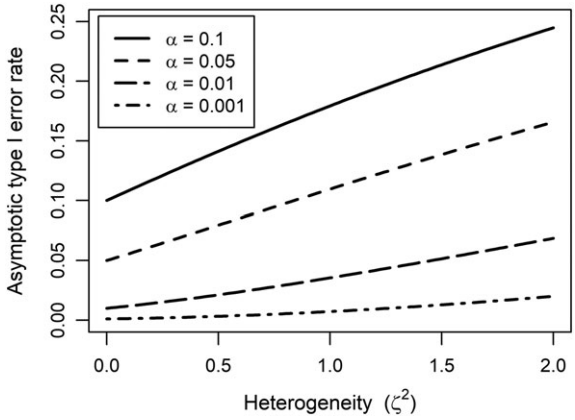

FIGURE 2 Inflated asymptotic type I error rate for the test of hypothesis $H_0 : \beta_F = 0$ in the presence of heterogeneity, when using a naive estimator of the variance from Equation 5 for a simple case of difference in means of continuous normal outcome with constant variance and balanced study designs (see details in Appendix E1)

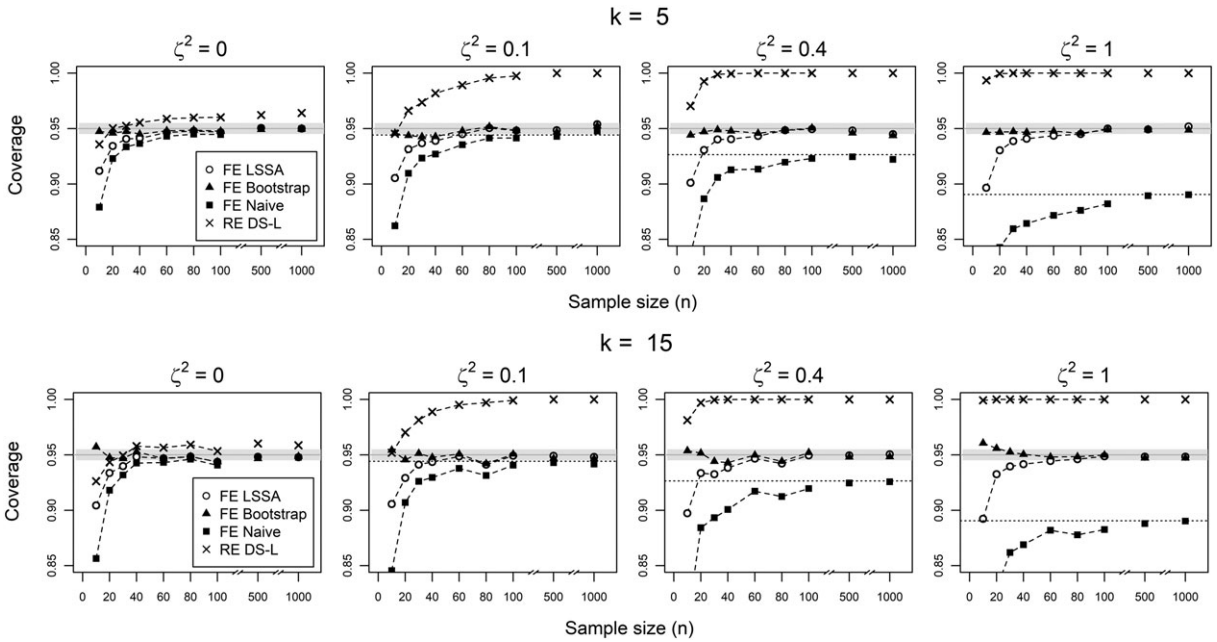

FIGURE 3 Coverage probabilities of 95% confidence intervals for $\beta_F = 0$ from 10 000 simulations, using the fixed effects approach large sample size approximation (LSSA) estimator for the variance and bootstrap percentiles, compared to the "naive" estimator from a common effect approach and the DerSimonian-Laird estimator from a random effects approach

the asymptotic variance of the information ϕ_i, implying that the inflation factor increases when the studies are less informative about ϕ. Figure 2 illustrates, for a simple case, the nontrivial impact of inflation on type I error rates when testing a point null hypothesis for β_F, even in large samples. The overstatements of statistical significance depend on the heterogeneity present, but also the nominal level α. (Full details are given in Appendices B and E1). This theoretic result underpinning Figure 2 has been confirmed empirically in a simulation study (Figure 3), described in Section 4.3.

3.3 | A parameter to quantify heterogeneity

While quantifying heterogeneity in meta-analyses has an obvious scientific appeal—describing how effects differ across study populations—the results of Section 3.2 do also suggest a statistical role for consideration of heterogeneity. Bridging these 2 goals, we now propose a parameter to quantify the heterogeneity of a group of effect-size parameters.

As a natural extension of the location-summary β_F, we define

$$\zeta^2 = \frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}(\beta_i - \beta_F)^2}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}} = \frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\beta_i^2}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}} - \beta_F^2 = \frac{\sum_{i=1}^{k} \eta_i \phi_i \beta_i^2}{\sum_{i=1}^{k} \eta_i \phi_i} - \beta_F^2, \tag{9}$$

where β_F is as in Equation 4. For fixed sample size proportions $\eta_1, \ldots, \eta_i$, we can see that ζ^2 is also a population parameter, just like β_F. We can interpret ζ^2 as a weighted average of the squared deviations of each study effect size from the weighted average effect β_F, where the weights are proportional to the precision (or the proportion of information) associated with each study effect. Consequently, deviations from more precisely estimable study effects are upweighted. This parameter ζ^2 is a weighted average squared deviation and quantifies the heterogeneity of the effect sizes.

As shown in Appendix C, ζ^2 can also be defined without regard to β_F, as a summary of pairwise comparisons of the β_i, by writing it in the form

$$\zeta^2 = \frac{1}{2}\frac{\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{1}{\sigma_i^2}\frac{1}{\sigma_j^2}(\beta_i - \beta_j)^2}{\left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right)^2} = \frac{\sum_{1 \le i < j \le k} \frac{1}{\sigma_i^2}\frac{1}{\sigma_j^2}(\beta_i - \beta_j)^2}{\left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right)^2}.$$

Specifically, ζ^2 is the weighted average of the pairwise differences of the effect sizes, weighting each pair by the product of their corresponding precisions. Unlike the between-studies variance τ^2 used in random effects approaches, ζ^2 is defined on just the studies at hand, not a hypothetical population of potential studies, and some scheme for sampling from this population.

Although the definition of ζ^2 is free of distributional assumptions, it can further justified if we assume normality of the effect-size estimators (see, eg, Rice et al[4]). Under this assumption, the Q statistic is distributed noncentral χ^2 with $k-1$ degrees of freedom and noncentrality parameter λ given by

$$\lambda = \sum_{i=1}^{k} \frac{1}{\sigma_i^2}(\beta_i - \beta_F)^2 = \left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right)\left(\frac{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}(\beta_i - \beta_F)^2}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}\right) = \left(\sum_{i=1}^{k} n_i \phi_i\right)\zeta^2 = \Phi\zeta^2, \tag{10}$$

where we have used $\Phi = \sum n_i \phi_i = N \sum \eta_i \phi_i$ to denote the total amount of information. This expression means that λ, and thus the power of the test of homogeneity based on Q, depends on 2 components: one is the total amount of information, which in turn depends on the total sample size, and the other is the heterogeneity between effect sizes, as given by ζ^2, which is independent of the total sample size. In other words, ζ^2 provides a measure of the distance from the null hypothesis of homogeneity.

4 | INFERENCE FOR β_F AND ζ^2

4.1 | Inference for β_F and ζ^2 with known standard errors

Inference for $\hat{\beta}_F$ with known standard errors was described in Equation 2; confidence intervals for β_F are usually built from a normal approximation, appealing to the large sample properties of $\hat{\beta}_i$. For a full description, see, eg, Hartung and Knapp.[6]

For the estimation of the heterogeneity parameter ζ^2, with known standard errors and efficient $\hat{\beta}_i$, we write $\sigma_i^2 = (n_i\phi_i)^{-1}$ for $i = 1, \ldots, k$ and also define $\Phi = \sum_{i=1}^{k} n_i\phi_i$ as the "total information." Then with no further distributional assumptions, a simple moment-based point estimate of ζ^2 is given by

$$\hat{\zeta}^2 = \frac{\sum_{i=1}^{k} \sigma_i^{-2}(\hat{\beta}_i - \hat{\beta}_F)^2 - (k-1)}{\sum_{i=1}^{k} \sigma_i^{-2}} = \frac{Q - (k-1)}{\Phi}, \tag{11}$$

with details given in Appendix D. To give a strictly positive estimator of ζ^2, we can report

$$\hat{\zeta}_0^2 = \max\left(0, \frac{Q - (k-1)}{\Phi}\right).$$

To obtain approximate confidence intervals for ζ^2, we assume normality of the effect-size estimators and exploit the relationship between ζ^2 and the noncentrality parameter λ as given in Equation 10. We proposed using methods for con-

structing exact confidence intervals for the noncentrality parameter of a chi-square distribution that have been proposed and evaluated previously.[42,43] Basically, these methods consist on inverting a probability interval of the non-central χ^2 distribution. For example, for given for given Φ and Q, a $(1-\alpha)\times 100\%$ confidence interval for ζ^2 is given by all the values for which

$$\chi^2_{k-1,\alpha/2}(\Phi\zeta^2) \leq Q \leq \chi^2_{k-1,1-\alpha/2}(\Phi\zeta^2).$$

Solutions can be obtained numerically, and code for this and other types of confidence intervals (for the noncentrality parameter) is available.[43]

4.2 | Inference on for β_F and ζ^2 with estimated standard errors

4.2.1 | Large sample size approximation

Based on Equation 8, a large sample size approximation (LSSA) of the variance of $\hat{\hat{\beta}}_F$ is given by

$$\text{Var}[\hat{\hat{\beta}}_F] \approx \frac{1}{N\left(\sum_i^k \eta_i\phi_i\right)}\left[1+\frac{\sum_i^k \eta_i(\beta_i-\beta_F)^2 f_i(\boldsymbol{\theta}_i)}{\sum_i^k \eta_i\phi_i}\right]. \tag{12}$$

Further details on the specific form of $f_i(\boldsymbol{\theta}_i)$ in (12) for some common effect-size estimators are provided in Appendix E. In situations where the function f_i is known or can be estimated, tests of hypothesis and confidence intervals can be based on a normal approximation using a plug-in estimator of (12) with the estimates of β_i, ϕ_i and $\boldsymbol{\theta}_i$ for $1 \leq i \leq k$ and β_F. The $(1-\alpha)\times 100\%$ LSSA interval then takes the form

$$\hat{\hat{\beta}}_F \pm z_{1-\alpha/2}\frac{1}{\sqrt{\sum_{i=1}^k n_i\hat{\phi}_i}}\left[1+\frac{\sum_i^k n_i\left(\hat{\beta}_i-\hat{\hat{\beta}}_F\right)^2 f_i(\hat{\boldsymbol{\theta}}_i)}{\sum_i^k n_i\hat{\phi}_i}\right]^{1/2}.$$

4.2.2 | Quasi-F approach

We next propose an interval based on inverting a test of the null hypothesis of homogeneity, similarly to Hartung and Knapp.[5] It is based on a "quasi-F" test statistic, a statistic that approximates a F-distributed random variable.[44]

To construct it, we use normality of the $\hat{\beta}_i$ to provide

$$\frac{(\hat{\beta}_F-\beta_{F0})^2}{\text{Var}[\hat{\beta}_F]} \sim \chi^2_1,$$

$$Q \sim \chi^2_{k-1}(\lambda), \text{ with } \lambda = \sum_{i=1}^k n_i\phi_i(\beta_i-\beta_F)^2 = \Phi\zeta^2,$$

for null value β_{F0}. Approximating the noncentral χ^2 distribution by matching its moments to a central χ^2,[45,46] we can approximate the distribution of Q as a α-scaled central χ^2 distribution with ν degrees of freedom ($\alpha\chi^2_\nu$), where

$$\alpha = 1+\frac{\lambda}{(k-1)+\lambda} = 1+\frac{\Phi\zeta^2}{(k-1)+\Phi\zeta^2}$$

$$\nu = (k-1)+\frac{\lambda^2}{(k-1)+2\lambda} = (k-1)+\frac{(\Phi\zeta^2)^2}{(k-1)+2\Phi\zeta^2}.$$

Under the assumptions above, Q and $\hat{\beta}_F$ are independent,[4,5] so

$$\frac{(\hat{\beta}_F-\beta_0)^2/\text{Var}[\hat{\beta}_F]}{Q/\alpha\nu}$$

has an approximate F^1_ν distribution, and its signed square root has an approximate Student t distribution with ν degrees of freedom.

To use these results with unknown σ_i, a "quasi-F" statistic can be constructing by plugging-in estimators of all those quantities. Thus, letting $\hat{\hat{\beta}}$ be as in Equation 5, $\widehat{\text{Var}}[\hat{\hat{\beta}}_F]$ the LSSA given in Equation 12, along with plug-in estimates of Q, ζ^2, and φ, used in turn to estimate α and ν. Taking square roots, the test statistic

$$t = \sqrt{\frac{\hat{\alpha}\hat{v}(\hat{\hat{\beta}}_F - \beta_{F0})^2}{\widehat{\text{Var}}[\hat{\hat{\beta}}_F]\hat{Q}}} \tag{13}$$

has an approximate Student t distribution with $\hat{v}$ degrees of freedom under the null hypothesis $H_0 : \beta_F = \beta_{F0}$. (This reference distribution would differ importantly from a standard normal for small values of $\hat{v}$, which would be expected when the meta-analysis includes few studies (small k) and the total amount of information times the amount of heterogeneity is small, ie, approaching the limit where $\varphi\zeta^2 \to 0$.) Inverting this test, we obtain an approximate confidence interval for β_F.

4.2.3 | Parametric bootstrap

The alternative estimators described in Sections 4.2.1 and 4.2.2, which take into account the potential heterogeneity of the effect-size parameters, are based on approximations that would be expected to work in large sample settings, but would probably perform poorly in settings with very small size samples. An alternative method that could better in small sample size settings is bootstrap re-sampling. As individual-level observations are typically not available, we consider using parametric bootstrap sampling. (For a full review of this approach, see chapter 6 of Efron and Tibshirani[47])

Estimates of the variance of $\hat{\hat{\beta}}_F$, as well as 95% confidence intervals, and/or P values for testing of hypothesis can all be obtained from parametric sampling, based on the estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ and $s_1^2, s_2^2, \ldots, s_k^2$. Assuming a normal distribution of the effect sizes estimates, a parametric bootstrap sample of size B for each of the effect-size parameters β_i can be obtained:

$$\hat{\beta}^*_{i[b]} \sim N(\hat{\beta}_i, s_i^2), \text{ for } i = 1, \ldots, k; b = 1, \ldots, B. \tag{14}$$

However, parametric sampling for the variances of the effect estimates depends on the specific variance estimator used in each study. For example, for the variance of the difference in means of independent groups where equal variances are assumed, a bootstrap sample of $\hat{\sigma}_i^2$ can be obtained as

$$\hat{\sigma}^{2*}_{i[b]} = \frac{\hat{\varsigma}^{*2}_{i[b]}}{n_i} \quad \text{with } \hat{\varsigma}^{2*}_{i[b]} \sim \frac{\hat{\varsigma}_i^{\,2}}{n_i - 2}\chi^2_{n_i-2}, \text{ for } i = 1, \ldots, k; b = 1, \ldots, B, \tag{15}$$

where $\hat{\varsigma}_i^2$ is the pooled estimate of the common variance ς_i^2.[48] More generally, for estimates from linear regression (where normality and constant variance are assumed), the sampling can be done from a χ^2 distribution with $(n_i - p_i)$ degrees of freedom, where p_i denotes the number of predictors in the regression (including the intercept). In contrast, when β_i is estimated as the difference in means of independent groups with the variances not assumed to be equal, the parametric sampling of $\varsigma^2_{i,X}$ and $\varsigma^2_{i,Y}$ should be done separately and then combined to obtain the value of σ_i^2. Further details on the specific form of some of these estimators can found in Appendix E.

From the parametric bootstrap samples of effect size and variance estimators, different estimates and/or test statistics can be obtained. We propose (and evaluate) the following:

1. A pivotal $(1-\alpha)$% confidence interval based on a normal approximation and using an estimate of the variance of $\hat{\hat{\beta}}_F$ from a bootstrap sample (see chapter 6 of Efron and Tibshirani[47]):

$$\hat{\hat{\beta}}_F \pm z_{1-\alpha/2}\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}^*_{F[b]} - \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^*_{F[b]}\right)^2}, \text{ where } \hat{\beta}^*_{F[b]} = \frac{\sum_{i=1}^{k}\frac{1}{\hat{\sigma}^{2*}_{i[b]}}\hat{\beta}^*_{i[b]}}{\sum_{i=1}^{k}\frac{1}{\hat{\sigma}^{2*}_{i[b]}}}$$

2. A $(1-\alpha)$% confidence interval constructed from the percentiles of the empirical distribution of the bootstrap sample of $\hat{\beta}^*_{F[b]}$, as defined in (1), (see chapter 13 of Efron and Tibshirani[47]):

$$\left(\hat{\beta}^*_{F(\alpha/2)}, \hat{\beta}^*_{F(1-\alpha/2)}\right).$$

3. A Bootstrap-t confidence interval (see chapter 12 of Efron and Tibshirani[47]), based on the percentiles from the distribution of a test statistic constructed using a "naive" estimator of the variance of $\hat{\hat{\beta}}_F$:

$$\left(\hat{\hat{\beta}}_F - t^*_{(1-\alpha/2)}\sqrt{\left(\sum_{i=1}^{k}\frac{1}{s_i^2}\right)^{-1}}, \hat{\hat{\beta}}_F - t^*_{(\alpha/2)}\sqrt{\left(\sum_{i=1}^{k}\frac{1}{s_i^2}\right)^{-1}}\right), \text{where } t^*_{[b]} = \frac{\hat{\beta}^*_{F[b]}}{\sqrt{\left(\sum_{i=1}^{k}\frac{1}{\hat{\sigma}^{2*}_{i[b]}}\right)^{-1}}}.$$

4. A Bootstrap-t confidence interval, based on the percentiles from the distribution of a test statistic constructed using the LSSA estimator of the variance of $\hat{\hat{\beta}}_F$, as given in (12):

$$\left(\hat{\hat{\beta}}_F - t^*_{(1-\alpha/2)}\sqrt{\widehat{\text{Var}}[\hat{\hat{\beta}}_F]}, \hat{\hat{\beta}}_F - t^*_{(\alpha/2)}\sqrt{\widehat{\text{Var}}[\hat{\hat{\beta}}_F]}\right), \text{where } t^*_{[b]} = \frac{\hat{\beta}^*_{F[b]}}{\sqrt{\text{Var}^*_{[b]}[\hat{\hat{\beta}}_F]}}.$$

Similar approaches are proposed for the heterogeneity parameter ζ^2, based on a bootstrap sample of the estimator proposed in (14) and (15)

$$\hat{\zeta}^{2*}_{[b]} = \frac{\sum_{i=1}^{k}\hat{\sigma}^{-2*}_{i[b]}(\hat{\beta}^*_{i[b]} - \hat{\beta}^*_{F[b]})^2 - (k-1)}{\sum_{i=1}^{k}\hat{\sigma}^{-2*}_{i[b]}}, \text{for } b = 1, \ldots, B.$$

We present evaluations of the coverage of confidence intervals using 2 approaches (some other alternatives were attempted, but did not show important improvement):

1. A pivotal $(1-\alpha)\%$ confidence interval based on a normal approximation and using an estimate of the variance of $\hat{\zeta}^2$ from the bootstrap sample:

$$\hat{\zeta}^2 \pm z_{1-\alpha/2}\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\zeta}^{2*}_{[b]} - \frac{1}{B}\sum_{b=1}^{B}\hat{\zeta}^{2*}_{[b]}\right)^2},$$

   where $\hat{\zeta}^{2*}_{[b]}$ is defined as above.
2. A $(1-\alpha)\%$ confidence interval constructed from the percentiles of the empirical distribution of the bootstrap sample of $\hat{\zeta}^{2*}_{[b]}$:

$$\left(\hat{\zeta}^{2*}_{(\alpha/2)}, \hat{\zeta}^{2*}_{(1-\alpha/2)}\right).$$

4.3 | Simulation study

We conducted a simulation study to evaluate and compare the different estimation methods proposed for β_F and ζ^2. For our simulations, we considered fixed effect sizes $(\beta_1, \ldots, \beta_k)$, uniformly spaced and centered around zero ($\beta_F = 0$), with the spacing in between given by fixed values of ζ^2. We assumed continuous normal outcomes and the effect size β_i given by the mean difference between 2 groups, assuming equal population variances and balanced designs. We took random draws of the effect estimates $(\hat{\beta}_1, \ldots, \hat{\beta}_k)$ from normal distributions centered around the fixed effects $(\beta_1, \ldots, \beta_k)$ along with random draws of their variances taken from scaled χ^2 distributions with $n_i - 2$ degrees of freedom. Various scenarios were considered, varying the number of studies, sample sizes, and amount of heterogeneity. In addition to the various confidence intervals proposed here for β_F and ζ^2, we also compared their performance with methods typically used in meta-analysis, ie, the common effect and random effects approaches. To aid the comparisons, we chose a setup in which all these approaches estimate location parameters with the same numerical value. Further details on the settings and complete results from the simulation study can be found in the supporting information for the online article.

Representative results are shown in Figure 3. For the estimation of the location parameter β_F, we observed a better performance of parametric bootstrap methods over those based on asymptotic approximations, especially with small sample sizes. Among these, the confidence interval based on the percentiles of the empirical distribution of the parametric sample would be recommended, because it is simple and performed well, providing coverage close to nominal level. However, we also notice that the LSSA method performed reasonably well for large sample sizes (at least 60 subjects per study) and note that it can be used if the parametric bootstrap could not be implemented.

Compared to existing methods, as expected, the random effects approach (using the DerSimonian-Laird estimator of μ) provided overconservative inference, as result of wide confidence intervals that account for random sampling of effect

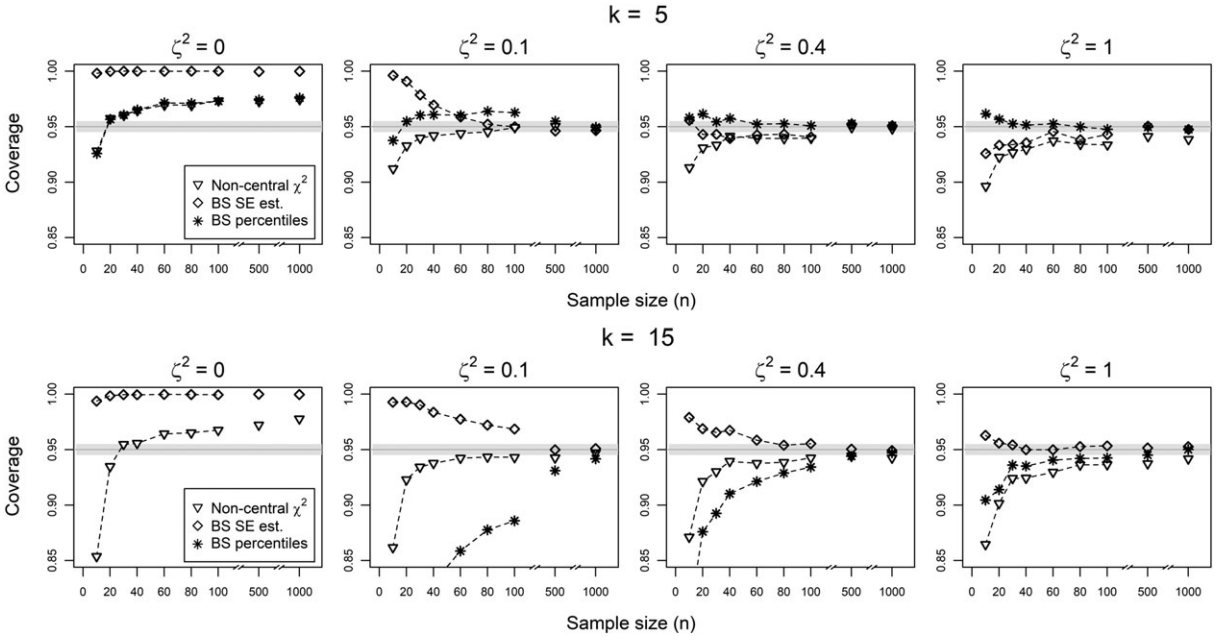

FIGURE 4 Coverage probabilities of 95% confidence intervals for ζ^2 from 10 000 simulations, using an inverted probability interval from a noncentral χ^2 distribution, a normal approximation with bootstrap estimate of the standard error and a bootstrap estimate based on the quartiles of the empirical distribution

sizes that is not present in our simulation settings. However, for the common effect estimator, which is equivalent to use a naïve estimate of the variance of β_F as given in Equation 7, the coverage probability approaches the nominal level as the sample size increases but never reaching it in the presence of heterogeneity. The asymptotic coverage of this naive estimator has been calculated using (12) and is shown as dotted horizontal lines in Figure 3 (see details in Appendix E1).

For the heterogeneity parameter ζ^2, although all the proposed methods seemed to asymptotically achieve the nominal coverage probability, none of them performed uniformly better for small sample size settings in all scenarios (Figure 4). The normal approximation with a moment based estimate of the standard error showed both significant overcoverage and undercoverage in different scenarios (not shown). The normal approximation using a Bootstrap estimate of the standard error seemed to correct the undercoverage in some scenarios, but not when the number of studies was small ($k = 3$), while the bootstrap confidence intervals based on the percentiles showed important undercoverage for low values of heterogeneity and large number of studies ($k = 7, 15$). This result is consistent with a previous result, in which the consistency of bootstrap estimation is related to the asymptotic normality of the statistic,[49,50] while in our case, distribution of the statistic is far from normal, for small sample size and low level of heterogeneity. On the other hand, given the more consistent performance of the inverted probability interval from a noncentral χ^2 distribution, we would recommend its use when the sample sizes are large enough (at least 40 observations per study) and the studies are not strongly heterogeneous.

5 | EXAMPLE

In this section, we apply the estimation methods discussed in Section 4 to an example from a systematic review of studies that evaluate the efficacy of zinc in reducing the incidence, severity, and duration of common cold symptoms.[51] In this particular meta-analysis, the authors included studies that compare zinc acetate lozenges with placebo, with the outcome being the duration of cold symptoms (in days) and the treatment effect measured by the mean difference. A forest plot is shown in Figure 5.

In Table 2, we summarize the results of meta-analyses on the 6 studies comparing zinc lozenges to placebo, using 3 different approaches. We observe that the point estimates of β_0 and β_F from the common effect and fixed effects approaches, respectively, although numerically the same ($\hat{\beta}_0 = \hat{\beta}_F = -2.04$ days), estimate different parameters. The first estimates a common effect underlying all 6 studies, but given the evident heterogeneity between studies, this inference does not seem to be adequate, or even valid. On the other hand, $\hat{\beta}_F$ estimates a weighted average of the mean differences from the 6 studies, for which a significant amount of heterogeneity is observed, as reflected by the estimate of ζ^2. More specifically, $\hat{\beta}_F$ estimates the mean difference in duration of common cold averaged in a meta-population composed of the populations from which the samples of these 6 studies were drawn, in proportions given by $\sigma_i^{-2} / \sum_i^k \sigma_i^{-2}$. Similarly, ζ^2 can be thought

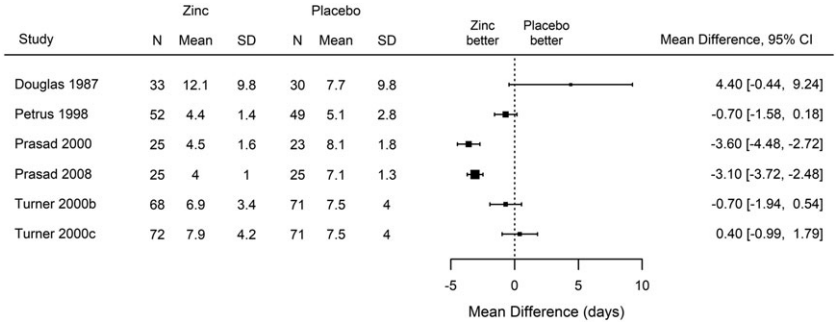

FIGURE 5 Meta-analysis on the efficacy of zinc acetate lozenges in reducing the duration of cold symptoms[51]

TABLE 2 Fixed and random effects approaches to the meta-analyses on the effect of zinc acetate lozenges, estimated as the mean difference in the duration of symptoms of the common cold (in days),[51] with point estimates and 95% confidence intervals obtained from different methods of estimation

Common effect	$\hat{\beta}_0$ (95% CI)	
	−2.04 (−2.45, −1.64)	
Fixed effects	$\hat{\beta}_F$ (95% CI)	$\hat{\zeta}^2$ (95% CI)
Assuming known σ_i		
Naive estimator	−2.04 (−2.45, −1.64)	
Noncentral χ^2 inverted test		2.09 (1.09, 3.50)
Unknown σ_i		
Large sample size approximation (LSSA)	−2.04 (−2.51, −1.57)	
Quasi-F–based Student t	−2.04 (−2.53, −1.55)	
Parametric bootstrap (B = 2000)	−2.04 (−2.54, −1.59)	2.09 (1.15, 3.68)
Random effects	$\hat{\mu}$ (95% CI)	$\hat{\tau}^2$ (95% CI)
DerSimonian-Laird	−1.21 (−2.69, 0.28)	2.81 (1.19, 46.0)
Maximum likelihood	−1.21 (−2.69, 0.28)	2.79 (0, 6.53)
Restricted maximum likelihood	−1.13 (−2.83, 0.57)	3.78 (0, 9.25)
Sidik-Jonkman	−1.02 (−3.06, 1.01)	5.66 (2.20, 34.04)

as estimating how far apart the mean differences in 2 of these populations are, averaged over the same meta-population. We also observe that the results from different estimation methods, although not exactly the same, do not seem to differ importantly, with a difference in length of 0.13 days between the 95% confidence intervals using the LSSA and the parametric bootstrap.

On the other hand, random effects meta-analysis estimates the mean and variance of a population from which the effects in the 6 studies are thought to have been drawn (μ and τ^2). The inference now is not made for the population of subjects (on whom we wish to estimate an average effect of a treatment) but for a population of potential treatment effects. As shown in Table 2, different methods for estimating the between-studies variance give notably different results, with larger estimates of τ^2 yielding estimates of μ that are closer to the unweighted simple average of the study effects (−0.56). Moreover, the precision with which these parameters are estimated is much smaller than the precision with which β_F and ζ^2 are estimated, even after taking into account the uncertainty in the estimation of the variances. This gain in precision, it should be noted, is not a result of a particular choice of estimation technique, it is instead the result of targeting our inference to a parameter that is easier to estimate, ie, one for which the the data provide most information.

To further illustrate the properties of the estimators of β_F and ζ^2 in a fixed effects meta-analysis, we have modified the example into 3 different versions, as shown in Figure 6. First, we increased the precisions of the estimates in the meta-analysis, by artificially growing the sample sizes by a factor of 10 (panel B). This results in a greater precision for the estimates of β and ζ^2. However, this same increase in information does not translate into an increased precision for estimating μ or τ^2 in a random-effects model (for which more studies, rather than larger sample sizes, would be needed). In another version of the meta-analysis, we have kept the same precision but shrunk (shifted) the estimates towards $\hat{\beta}_F$, so that the squared deviations have been reduced by factor of 10 (panel C). Reflecting this relative homogeneity, the estimate of ζ^2 is much lower and close to zero. We also notice that estimate of β remains practically unchanged, ie, is mostly independent of $\hat{\zeta}^2$ (except for the variance inflation effect described in Section 3.2, which is not substantial in this case).

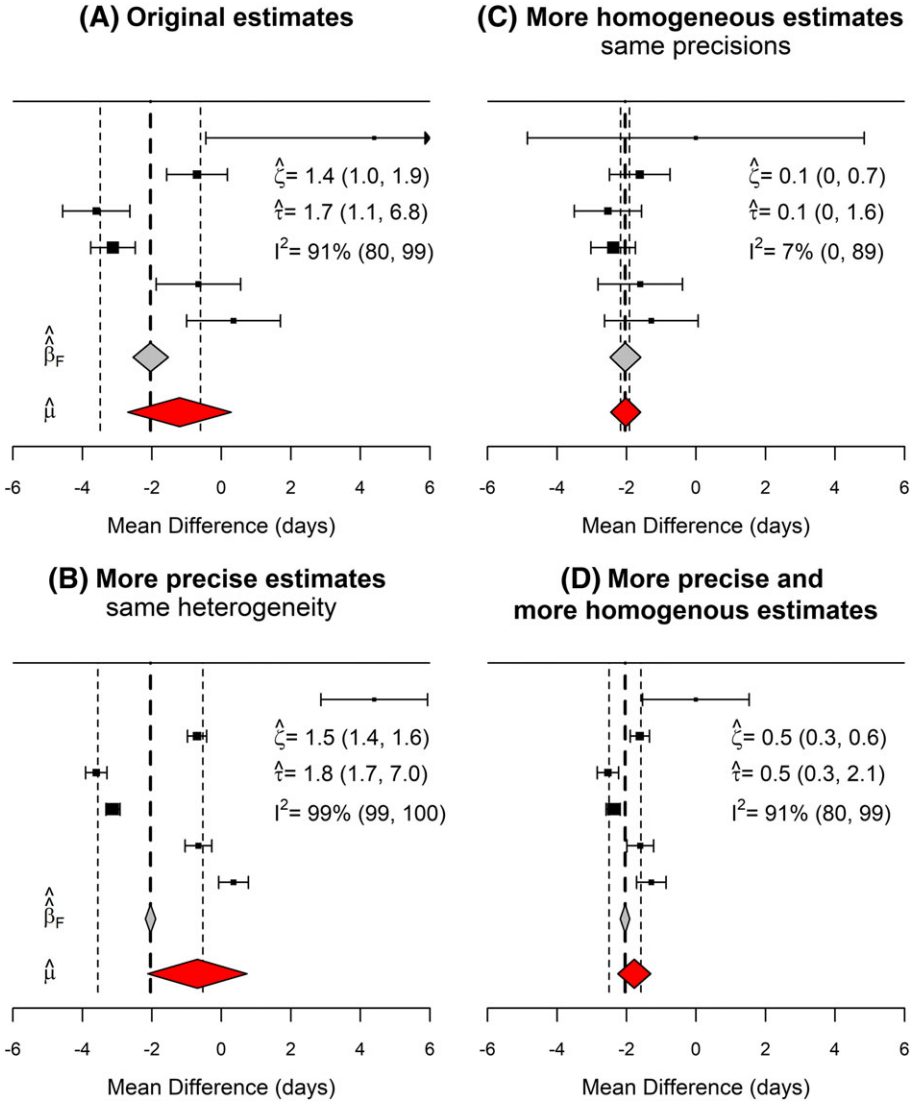

FIGURE 6 Location and scale parameter estimates for 4 different versions of the meta-analysis in Figure 5: A, original estimates; B, more homogeneous estimates with squared deviation from $\hat{\beta}_F$ reduced by a factor of 10; C, more precise estimates with sample sizes 10 times those of the original estimates; D, more precise and more homogeneous estimates, with sample sizes 10 times larger and squared deviations from $\hat{\beta}_F$ reduced by a factor of 10, relative to the original estimates. Rhomboids are used to represent point estimates and 95% confidence intervals of location parameters β_F (in gray) and μ (in red). The vertical dashed lines represent the square root of the estimated averaged squared deviations from β_F, as given by $\hat{\zeta} = \sqrt{\hat{\zeta}^2}$

In contrast, the estimate of μ, both in terms of its location and precision, is highly dependent on between study variance, as estimated by $\hat{\tau}^2$. Lastly, we artificially reduced the between-study heterogeneity and the within-study variance by the same factor (panel D). As a result of this, the value of the Q statistic is exactly the same as in the original version of the meta-analysis ($Q = 53.8$, 5 degrees of freedom, P value $< .0001$), and so is the value for I^2. This makes sense, as in both meta-analyses, heterogeneity accounts for the same proportion of the total variation. However, in absolute terms, the estimates in the modified version are much closer to each other than in the original meta-analysis, and this is picked up by estimates of ζ^2 and τ^2, as they are both quantify "absolute" heterogeneity. Their confidence intervals in both cases exclude zero, rejecting the null hypothesis of homogeneous effects. However, as pointed out before, we can estimate ζ^2 with higher precision, even with few studies.

6 | DISCUSSION

In this paper, we have addressed several aspects of the fixed effects meta-analysis with within-study estimates of the standard errors. To formally motivate its precision weighting, we described the optimality of the corresponding parameter

β_F, and by studying the behavior of the precision-weighted estimate in detail, we showed the important role of a particular measure of heterogeneity, ζ^2.

Frequentist methods for the estimation of both the location parameter β_F and the heterogeneity parameter ζ^2 were proposed, including corrected estimators that take into account the uncertainty in the estimation of the within study variances. Estimation methods based on asymptotic approximations, as well as methods based on parametric bootstrap, were implemented and have been evaluated in a simulation study.

In the results of our simulation study, we observed a better performance of parametric bootstrap methods over those based on asymptotic approximations for the estimation of the location parameter β_F, specially in small sample size settings. Among these, the confidence interval based on the percentiles of the empirical distribution of the parametric sample would be recommended, because of its simplicity and good performance. However, we also notice that the LSSA method performed reasonably well for large sample sizes ($n \geq 60$, per study) and could be used if the parametric bootstrap can not be implemented.

For the heterogeneity parameter ζ^2, although no method performed uniformly better, the construction of 95% confidence intervals by inverting the probability interval from a noncentral χ^2 distribution seems to provide close to nominal coverage when the sample size is large enough (around 40 observations per study).

The main limitation in our simulation study is that the proposed methods were implemented with knowledge of how the study estimates (including standard errors) were generated. The independence of the point estimate and standard errors—plausible in most uses of linear regression—may not be as realistic if the study-specific analyses use logistic regression, or other forms of analysis under strong mean-variance relationships. The normality of the $\hat{\beta}_i$ may also be considered a limitation, but unless the outcome variable is very heavy-tailed and/or sensitive to a few observations, standard central limit theorem arguments suggest that this will only be an issue in extremely small samples.

We also illustrated the results of different estimation methods, as well as different approaches, with a previously published meta-analysis. This example, along with the results of our simulation study, supports the idea of approaching meta-analysis under a fixed effects framework, as a valid alternative to the typically used common effect and random effect approaches. Our approach, based on the estimation of both a location and a heterogeneity parameter, is more flexible than the restrictive common effect approach while allowing inference on the population of interest. Our approach also makes it unnecessary to choose between statistical models based on their adequacy rather than the target inference.

Finally, although we believe that estimation of both β_F and ζ^2 is useful for describing and combining in a meaningful way the effects of studies included in a meta-analysis, we propose their estimation only as part of a full battery of qualitative and quantitative tools that should be used to review, summarize, and synthesize a group of studies. No single parameter or estimator can always appropriately summarize all there is to say in a systematic review of medical studies, and practitioners should be encouraged and helped to understand the measures they choose to provide.

ACKNOWLEDGMENTS

Clara Dominguez-Islas gratefully acknowledges support from the University of Washington Department of Biostatistics and the MRC Biostatistics Unit (Programme number U105260558).

ORCID

Clara Dominguez Islas http://orcid.org/0000-0003-0653-3441

REFERENCES

1. Glass GV. Primary, secondary, and meta-analysis of research. *Educ Res*. 1976;5(10):3-8.
2. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions Version 5.1.0 [updated March 2011]. *The Cochrane Collaboration*. 2011. Available from http://www.handbook.cochrane.org.
3. Cochrane. *Our Evidence*. http://www.cochrane.org/evidence. Accessed February 1, 2017.
4. Rice K, Higgins JPT, Lumley T. A re-evaluation of fixed effect(s) meta-analysis. *J. R. Stat. Soc. A*. 2018;181:205-227.
5. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Stat Med*. 2001;20(12):1771-1782.
6. Hartung J, Knapp G, Sinha BK. *Statistical Meta-Analysis with Applications*. Hoboken, NJ: John Wiley & Sons; 2008.
7. Borenstein M, Hedges LV, Higgins J, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*. 2010;1(2):97-111.

8. Thompson SG, Pocock SJ. Can meta-analyses be trusted? *The Lancet*. 1991;338(8775):1127-1130.
9. Gelman A, O'rourke K. Discussion: difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics*. 2014;15(1):18-23.
10. Peto R. Why do we need systematic overviews of randomized trials? (Transcript of an oral presentation, modified by the editors). *Stat Med*. 1987;6(3):233-240.
11. Laird NM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care*. 1990;6(1):5-30.
12. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.
13. Berrington de González A, Cox DR. Interpretation of interaction: a review. *Ann Appl Stat*. 2007;1(2):371-385.
14. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.
15. Higgins J, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc*. 2009;172(1):137-159.
16. Peto R. Discussion. *Stat Med*. 1987;6(3):241-244.
17. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. New York: Academic Press; 1985.
18. Lin DY, Zeng D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*. 2010;97(2):321-332.
19. Doob JL. Application of the theory of martingales. *Le calcul des probabilites et ses applications*. 1949:23-27.
20. Vaart AW. *Asymptotic Statistics*. Cambridge, United Kingdom: Cambridge University Press; 2000.
21. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med*. 2005;24(15):2401-2428.
22. Senn S. Trying to be precise about vagueness. *Stat Med*. 2007;26(7):1417-1430.
23. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal*. 2006;1(3):515-534.
24. Goel PK. Information measures and Bayesian hierarchical models. *J Am Stat Assoc*. 1983;78:408-410.
25. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester, United Kingdom: John Wiley & Sons; 2004.
26. Rücker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Stat Med*. 2009;28(5):721-738.
27. Fisher RA. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*. 1921;1:3-32.
28. Li Y, Shi L, Daniel RH. The bias of the commonly-used estimate of variance in meta-analysis. *Commun Stat Theory Methods*. 1994;23(4):1063-1085.
29. Böckenhoff A, Hartung J. Some corrections of the significance level in meta-analysis. Technical Report, SFB 475, Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund; 1998. http://hdl.handle.net/10419/77210. Accessed November 15, 2017.
30. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med*. 1998;17(8):841-856.
31. Higgins J, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558.
32. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ: Br Med J*. 2003;327(7414):557-560.
33. Hedges LV. A random effects model for effect sizes. *Psychol Bull*. 1983;93(2):388-395.
34. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educ Behav Stat*. 2005;30(3):261-293.
35. Rao PSRS, Kaplan J, Cochran WG. Estimators for the one-way random effects model with unequal error variances. *J Am Stat Assoc*. 1981;76(373):89-97.
36. Hardy RJ, Thompson SG. A likelinood approach to meta-analysis with random effects. *Stat Med*. 1996;15(6):619-629.
37. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. *Biometrika*. 1971;58(3):545-554.
38. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc*. 1977;72(358):320-338.
39. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *J R Stat Soc Ser C Appl Stat*. 2005;54(2):367-384.
40. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med*. 2007;26(1):37-52.
41. Konstantopoulos S, Hedges LV. Analyzing effect sizes: fixed-effects Models. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. New York: Russel Sage Foundation; 2009.
42. Kent JT, Hainsworth TJ. Confidence intervals for the noncentral chi-squared distribution. *J Stat Plan Inference*. 1995;46(2):147-159.
43. Kent JT. Calculations for the noncentral chi distribution. Technical report: The University of Leeds. Availble at http://www1.maths.leeds.ac.uk/~john/software/index.html; 2005. Accessed November 15, 2017.
44. Raudenbush SW. Analyzing effect sizes: random-effects models. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. New York: Russell Sage Foundation; 2009.
45. Patnaik PB. The Non-Central χ^2- and F-Distributions and their applications. *Biometrika*. 1949;36(1-2):202-232.
46. Hedges LV, Pigott TD. The power of statistical tests in meta-analysis. *Psychol Methods*. 2001;6(3):203-217.
47. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press; 1994.

48. Borenstein M. Effect sizes for continuous data. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. New York: Russell Sage Foundation; 2009.
49. Horowitz JL. The Bootstrap. In: Heckman JJ, Leamer E, eds. *Handbook of econometrics*. Amsterdam, The Netherlands: North-Holland/Elsevier; 2001.
50. Mammen E. *When Does Bootstrap Work?* New York: Springer; 1992.
51. Singh M, Das RR. Zinc for the common cold. *Cochrane Database Syst Rev*. 2013;(6):CD001364.https://doi.org/10.1002/14651858.CD001364.pub4.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Domínguez Islas C, Rice KM. Addressing the estimation of standard errors in fixed effects meta-analysis. *Statistics in Medicine*. 2018;37:1788–1809. https://doi.org/10.1002/sim.7625

APPENDIX A: PROOF OF LEMMA 1

Proof. Let $\boldsymbol{v}^T = (v_1, v_2, \ldots, v_k)$ be a vector of arbitrary weights with $\sum_{i=1}^k v_i = 1$, and let $\boldsymbol{v}^T\hat{\boldsymbol{\beta}} = \sum v_i\hat{\beta}_i$ be the estimator of $\boldsymbol{v}^T\boldsymbol{\beta} = \sum v_i\beta_i$, with $\text{Cov}(\hat{\boldsymbol{\beta}}) = \Sigma$, then

$$\text{Var}(\boldsymbol{v}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{v}^T\Sigma\boldsymbol{v}.$$

To minimize this expression, we use Lagrange multipliers:

$$\frac{d}{d\boldsymbol{v}}\left[\boldsymbol{v}^T\Sigma\boldsymbol{v} + \lambda(1 - \boldsymbol{v}^T\mathbf{1})\right] = 2\Sigma\boldsymbol{v} - \lambda\mathbf{1}$$
$$\frac{d}{d\lambda}\left[\boldsymbol{v}^T\Sigma\boldsymbol{v} + \lambda(1 - \boldsymbol{v}^T\mathbf{1})\right] = 1 - \boldsymbol{v}^T\mathbf{1}$$
$$\Rightarrow \boldsymbol{v} = \frac{\lambda}{2}\Sigma^{-1}\mathbf{1}; \quad \lambda = \frac{2}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}$$

so then

$$\boldsymbol{w} = \underset{\boldsymbol{v}:\boldsymbol{v}^T\mathbf{1}=1}{\text{argmin}}\left[\boldsymbol{v}^T\Sigma\boldsymbol{v}\right] = \frac{\Sigma^{-1}\mathbf{1}_k}{\mathbf{1}_k^T\Sigma^{-1}\mathbf{1}_k}$$

with

$$\text{Var}(\boldsymbol{w}^T\hat{\boldsymbol{\beta}}) = \frac{\mathbf{1}_k^T\Sigma^{-1}\Sigma\Sigma^{-1}\mathbf{1}_k}{(\mathbf{1}_k^T\Sigma^{-1}\mathbf{1}_k)^2} = \frac{1}{\mathbf{1}_k^T\Sigma^{-1}\mathbf{1}_k}.$$

□

APPENDIX B: DERIVATION OF THE ASYMPTOTIC VARIANCE OF $\hat{\hat{\beta}}_F$ (SECTION 3.2)

First, we write $s_i^2 = (n_i\phi_i)^{-1}$ as the estimator of $\sigma_i^2 = (n_i\phi_i)^{-1}$, the variance of $\hat{\beta}_i$ in the i^{th} study. We start by assuming that $\hat{\beta}_i$ and $\hat{\phi}_i$ are independent and have an asymptotic normal distribution:

$$\sqrt{n_i}\left(\begin{pmatrix}\hat{\beta}_i\\ \hat{\phi}_i\end{pmatrix} - \begin{pmatrix}\beta_i\\ \phi_i\end{pmatrix}\right) \xrightarrow{\text{d}} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\phi_i^{-1} & 0\\ 0 & f_i(\boldsymbol{\theta}_i)\end{pmatrix}\right), \tag{B1}$$

where $f_i(\boldsymbol{\theta}_i)$ is some function of the distributional moments of the population(s) in study i. For a meta-analysis of studies with different sample sizes, we define $\eta_i = n_i/N$, with $N = \sum_i^k n_i$. Then, dividing (B1) by $\sqrt{\eta_i}$, we get

$$\sqrt{N}\left(\begin{pmatrix}\hat{\beta}_i\\ \hat{\phi}_i\end{pmatrix} - \begin{pmatrix}\beta_i\\ \phi_i\end{pmatrix}\right) \xrightarrow{\text{d}} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\frac{1}{\eta_i\phi_i} & 0\\ 0 & \frac{f_i(\boldsymbol{\theta}_i)}{\eta_i}\end{pmatrix}\right).$$

Assuming that the study estimates are all independent, we can write

$$\sqrt{N}\left(\begin{pmatrix}\hat{\beta}_1\\ \vdots\\ \hat{\beta}_k\\ \hat{\phi}_1^2\\ \vdots\\ \hat{\phi}_k^2\end{pmatrix}-\begin{pmatrix}\beta_1\\ \vdots\\ \beta_k\\ \phi_1^2\\ \vdots\\ \phi_k^2\end{pmatrix}\right)\xrightarrow{d}N_{2k}\left(\begin{pmatrix}\mathbf{0}_k\\ \mathbf{0}_k\end{pmatrix},\begin{pmatrix}\text{Diag}_k\{1/\eta_i\phi_i\} & \mathbf{0}_{kk}\\ \mathbf{0}_{kk} & \text{Diag}_k\{f_i(\boldsymbol{\theta}_i)/\eta_i\}\end{pmatrix}\right). \tag{B2}$$

Recalling the definition of β_F:

$$\beta_F = g(\beta_1, \ldots \beta_k, \phi_1, \ \ldots \ , \phi_k) = \frac{\sum_i^k \eta_i\phi_i\beta_i}{\sum_i^k \eta_i\phi_i},$$

we obtain the following derivatives:

$$\frac{d}{d\beta_i}\beta_F = \frac{\eta_i\phi_i}{\sum_{i=1}^k \eta_i\phi_i}$$
$$\frac{d}{d\phi_i}\beta_F = \frac{\left(\sum_{i=1}^k \eta_i\phi_i\right)\eta_i\beta_i - \left(\sum_{i=1}^k \eta_i\phi_i\beta_i\right)\eta_i}{\left(\sum_{i=1}^k \eta_i\phi_i\right)^2} = \frac{\eta_i(\beta_i-\beta_F)}{\sum_{i=1}^k \eta_i\phi_i}.$$

Then, as long as $f_i(\boldsymbol{\theta}_i) < \infty$ for $i = 1, \ \ldots \ , k$, we can apply the delta method to B2, obtaining

$$\sqrt{N}\left[\left(\frac{\sum_i^k \eta_i\hat{\phi}_i\hat{\beta}_i}{\sum_i^k \eta_i\hat{\phi}_i}\right)-\left(\frac{\sum_i^k \eta_i\phi_i\beta_i}{\sum_i^k \eta_i\phi_i}\right)\right]\xrightarrow{d}N\left(0, \frac{1}{\sum_i^k \eta_i\phi_i}+\frac{\sum_i^k \eta_i(\beta_i-\beta_F)^2 f_i(\boldsymbol{\theta}_i)}{\left(\sum_i^k \eta_i\phi_i\right)^2}\right),$$

equivalently

$$\sqrt{N}\left(\hat{\hat{\beta}}_F-\beta_F\right)\xrightarrow{d}N\left(0, \frac{1}{\left(\sum_i^k \eta_i\phi_i\right)}\left[1+\frac{\sum_i^k \eta_i(\beta_i-\beta_F)^2 f_i(\boldsymbol{\theta}_i)}{\sum_i^k \eta_i\phi_i}\right]\right). \tag{B3}$$

Here, we notice that for some special cases when $f_i(\boldsymbol{\theta}_i)/\phi_i = c$, a constant, for $i = 1, \ \ldots \ , k$, this expression can be factorized out and the inflation factor can be then expressed as $(1 + c\zeta^2)$, a function of the heterogeneity parameter ζ^2 defined in Section 3.3. The specific form of $f_i(\boldsymbol{\theta}_i)$ for some common effect estimators of continuous outcomes are provided in Appendix E.

APPENDIX C: ALTERNATIVE EXPRESSION FOR ζ^2 (SECTION 3.3)

To simplify calculations, we write $w_i = \sigma_i^{-2}/\sum_{i=1}^k \sigma_i^{-2}$. Using this notation, we notice that $\sum_{i=1}^k w_i = 1$ and $\sum_{i=1}^k w_i\beta_i = \beta_F$. The parameter ζ^2 can then be written as

$$\begin{aligned}
\zeta^2 &= \sum_{i=1}^k w_i(\beta_i-\beta_F)^2 = \sum_{i=1}^k w_i\beta_i^2 - 2\sum_{i=1}^k w_i\beta_i\beta_F + \beta_F^2\\
&= \sum_{i=1}^k w_i\beta_i^2 - \sum_{i=1}^k w_i\beta_i\beta_F\\
&= \sum_{i=1}^k w_i\left(\sum_{j=1}^k w_j\right)\beta_i^2 - \sum_{i=1}^k w_i\beta_i\left(\sum_{j=1}^k w_j\beta_j\right)\\
&= \frac{1}{2}\sum_{i=1}^k\sum_{j=1}^k w_iw_j\beta_i^2 + \frac{1}{2}\sum_{i=1}^k\sum_{j=1}^k w_iw_j\beta_j^2 - \sum_{i=1}^k\sum_{j=1}^k w_iw_j\beta_i\beta_j\\
&= \frac{1}{2}\sum_{i=1}^k\sum_{j=1}^k w_iw_j(\beta_i-\beta_j)^2.
\end{aligned}$$

APPENDIX D: MOMENT BASED ESTIMATOR OF ζ^2 (SECTION 4.1)

Assuming known variances $\sigma_1^2, \ldots, \sigma_k^2$, with $\sigma_i^2 = (n_i\phi_i)^{-1}$ for $i = 1, \ldots, k$, we start by calculating the expected value of a plug-in estimate of ζ^2:

$$\begin{aligned}
\mathrm{E}\left[\frac{\sum_{i=1}^k n_i\phi_i(\hat{\beta}_i - \hat{\beta}_F)^2}{\sum_{i=1}^k n_i\phi_i}\right] &= \mathrm{E}\left[\sum_{i=1}^k \frac{n_i\phi_i}{\Phi}\hat{\beta}_i^2 - \hat{\beta}_F^2\right] \\
&= \sum_{i=1}^k \frac{n_i\phi_i}{\Phi}\left(\mathrm{Var}[\hat{\beta}_i] + (\mathrm{E}[\hat{\beta}_i])^2\right) - \left(\mathrm{Var}[\hat{\beta}_F] + (\mathrm{E}[\hat{\beta}_F])^2\right) \\
&= \left(\sum_{i=1}^k \frac{n_i\phi_i}{\Phi}\beta_i^2 - \beta_F^2\right) + \left(\frac{1}{\Phi}\sum_{i=1}^k n_i\phi_i\sigma_i^2 - \frac{1}{\Phi}\right) \\
&= \sum_{i=1}^k \frac{n_i\phi_i}{\Phi}(\beta_i - \beta_F)^2 + \frac{k-1}{\Phi} \\
&= \zeta^2 + \frac{k-1}{\Phi}.
\end{aligned}$$

Thus, an unbiased estimator of ζ^2 is then given by

$$\hat{\zeta}^2 = \frac{\sum_{i=1}^k \sigma_i^{-2}(\hat{\beta}_i - \hat{\beta}_F)^2 - (k-1)}{\sum_{i=1}^k \sigma_i^{-2}} = \frac{Q - (k-1)}{\Phi}.$$

APPENDIX E: ASYMPTOTIC VARIANCE OF ϕ FOR SOME COMMON EFFECT ESTIMATORS

In this section, we derive the asymptotic variance of the information parameter ϕ for some common estimators of treatment effect for continuous outcomes. This asymptotic variance, denoted $f(\theta_i)$, can then be plugged in estimators for the variance of $\hat{\hat{\beta}}_F$ described in Section 4.2 of the main paper.

E.1 | Asymptotic variance of ϕ for the mean difference of independent groups

Following Borenstein,[48] we first look at meta-analyses of studies that compare the means of 2 independent groups, when an assumption of equal variances is made. Here, the effect size in the i*th* study, $\beta_i = \Delta_i = \mu_{i,X} - \mu_{i,Y}$, is estimated by $\hat{\beta}_i = \bar{X}_i - \bar{Y}_i$, with $\mathrm{Var}(\hat{\beta}_i) = \sigma_i^2 = (1/n_{i,X} + 1/n_{i,Y})\varsigma_i^2$, where ς_i^2 is the population variance, assumed to be the same for the 2 groups in study i, and $n_{i,X}$ and $n_{i,Y}$ are the respective sample sizes (with $n_i = n_{i,X} + n_{i,Y}$). Here, we can write $\sigma_i^2 = (n_i\phi_i)^{-1}$ and $s_i^2 = (n_i\hat{\phi}_i)^{-1}$, with

$$\phi_i = g(\varsigma_i^2) = \left[\left(\frac{1}{n_{i,X}/n_i} + \frac{1}{n_{i,Y}/n_i}\right)\varsigma_i^2\right]^{-1} \tag{E1}$$

and $\hat{\phi} = g(\hat{\varsigma}_i^2)$, where $\hat{\varsigma}_i^2$ is the pooled estimator of the variance, given by

$$\hat{\varsigma}_i^2 = \frac{\sum_{j=1}^{n_{i,X}}(X_{ij} - \bar{X}_i)^2 + \sum_{j=1}^{n_{i,Y}}(Y_{ij} - \bar{Y}_i)^2}{n_i - 2} = \frac{(n_{i,X}-1)\hat{\varsigma}_{i,X}^2 + (n_{i,Y}-1)\hat{\varsigma}_{i,Y}^2}{n_{i,X} + n_{i,Y} - 2},$$

and $\hat{\varsigma}_{i,X}^2$ and $\hat{\varsigma}_{i,Y}^2$ are the sample variances of the 2 groups in study i. To obtain an asymptotic distribution for $\hat{\phi}_i$, we start from a standard result for the sample variance:

$$\sqrt{n_i}(\hat{\varsigma}_i^2 - \varsigma_i^2) \to_d N(0, (\kappa_i - 1)\varsigma_i^4),$$

where κ_i denotes to the kurtosis in the population distribution (which we will also assume to be the same in the 2 groups for now). So then, keeping the sample size proportions ($n_{i,X}/n_i$ and $n_{i,Y}/n_i$) fixed, the derivative of (E1) is

$$\frac{d}{d\varsigma_i^2} g(\varsigma_i^2) = -\left(\frac{1}{n_{i,X}/n_i} + \frac{1}{n_{i,Y}/n_i}\right)^{-1} (\varsigma_i^2)^{-2}.$$

The, applying the delta method, we conclude that

$$\sqrt{n_i}(\hat{\phi}_i - \phi_i) \rightarrow_d N\left(0, \frac{\kappa_i - 1}{\left(\frac{1}{n_{i,X}/n_i} + \frac{1}{n_{i,Y}/n_i}\right)^2 \varsigma_i^4}\right). \tag{E2}$$

We notice that for studies with balanced design ($n_{i,X} = n_{i,Y}$), the information $\phi_i = 1/4\varsigma_i^2$ and the asymptotic variance in the last expression reduces to $(\kappa_i - 1)/4^2\varsigma_i^4$. If all studies in a meta-analysis are balanced and the population variance and kurtosis can be assumed constant across all studies, then the inflation factor in (E2) is given by $(1 + \frac{\kappa-1}{4\varsigma^2}\zeta^2)$. This is the expression used to estimate the inflation in type I error rate, produced when the estimation of standard errors is not taken into account, illustrated in Figure 2 of the main paper.

Now, when the assumption of equal variances is not made, the variance of $\hat{\beta}_i = \bar{X} - \bar{Y}$ is given by $\text{Var}(\hat{\beta}_i) = \sigma_i^2 = \varsigma_{i,X}^2/n_{iX} + \varsigma_{i,Y}^2/n_{i,Y}$, where $\varsigma_{i,X}^2$ and $\varsigma_{i,Y}^2$ are the population variances of the 2 groups in the i*th* study.[48] Similar to the case of equal variances, here, we can also write $\sigma_i^2 = (n_i\phi_i)^{-1}$ and $s_i^2 = (n_i\hat{\phi}_i)^{-1}$, now with

$$\phi_i = g(\varsigma_{i,X}^2, \varsigma_{i,X}^2) = \left(\frac{\varsigma_{i,X}^2}{n_{i,X}/n_i} + \frac{\varsigma_{i,Y}^2}{n_{i,Y}/n_i}\right)^{-1},$$

and $\hat{\phi}_i = g(\hat{\varsigma}_{i,X}^2, \hat{\varsigma}_{i,X}^2)$. The asymptotic distribution for $\hat{\varsigma}_{i,X}^2$:

$$\sqrt{n_{i,X}}(\hat{\varsigma}_{i,X}^2 - \varsigma_{i,X}^2) \rightarrow_d N(0, (\kappa_{i,X} - 1)\varsigma_{i,X}^4),$$

can be expressed as

$$\sqrt{n_i}(\hat{\varsigma}_{i,X}^2 - \varsigma_{i,X}^2) \rightarrow_d N\left(0, \frac{(\kappa_{i,X} - 1)\varsigma_{i,X}^4}{n_{i,X}/n_i}\right),$$

and similarly for $\sqrt{n_i}(\hat{\varsigma}_{i,Y}^2 - \varsigma_{i,Y}^2)$. Keeping the sample size proportions within each study ($n_{i,X}/n_i$ and $n_{i,Y}/n_i$) fixed, we can write

$$\sqrt{n_i}\left[\begin{pmatrix}\hat{\varsigma}_{i,X}^2 \\ \hat{\varsigma}_{i,Y}^2\end{pmatrix} - \begin{pmatrix}\varsigma_{i,X}^2 \\ \varsigma_{i,Y}^2\end{pmatrix}\right] \rightarrow_d N_2\left[\mathbf{0}_2, \begin{pmatrix}\frac{(\kappa_{i,X}-1)\varsigma_{i,X}^4}{n_{i,X}/n_i} & 0 \\ 0 & \frac{(\kappa_{i,Y}-1)\varsigma_{i,Y}^4}{n_{i,Y}/n_i}\end{pmatrix}\right].$$

Taking derivatives, we find that

$$\nabla g(\varsigma_{i,X}^2, \varsigma_{i,X}^2) = -\left(\frac{\varsigma_{i,X}^2}{n_{i,X}/n_i} + \frac{\varsigma_{i,Y}^2}{n_{i,Y}/n_i}\right)^{-2} \left(\frac{1}{n_{i,X}/n_i}, \frac{1}{n_{i,Y}/n_i}\right)^T,$$

and applying the delta method, we obtain

$$\sqrt{n_i}(\hat{\phi}_i - \phi_i) \rightarrow_d N\left[0, \frac{\frac{(\kappa_{i,X}-1)\varsigma_{i,X}^4}{(n_{i,X}/n_i)^3} + \frac{(\kappa_{i,Y}-1)\varsigma_{i,Y}^4}{(n_{i,Y}/n_i)^3}}{\left(\frac{\varsigma_{i,X}^2}{n_{i,X}/n_i} + \frac{\varsigma_{i,Y}^2}{n_{i,Y}/n_i}\right)^4}\right].$$

E.2 | LSSA for the mean difference between 2 matched samples

When the effect size in each study is the mean difference between 2 matched samples, we have that $\hat{\beta}_i = \bar{X}_i - \bar{Y}_i$, with $\text{Var}(\hat{\beta}_i) = \sigma_i^2 = (\varsigma_{i,X}^2 + \varsigma_{i,Y}^2 - 2\rho\varsigma_{i,X}\varsigma_{i,Y})/m_i$, where ρ denotes the population correlation between any 2 matched observations X_{ij} and Y_{ij} in study i and $m_i = n_i/2$ is the number of the paired observations.[48] Assuming a bivariate normal distribution

for the observations (X_{ij}, Y_{ij}), the following asymptotic distribution can be obtained for the sample variances $(\hat{\varsigma}^2_{i,X}, \hat{\varsigma}^2_{i,X})$ and sample covariance $(\hat{\varsigma}_{i,XY} = \frac{1}{m_i}\sum_{j=1}^{m_i}(X_{ij} - \bar{X}_i)(Y_{ij} - \bar{Y}_i))$:

$$\sqrt{m_i}\left[\begin{pmatrix}\hat{\varsigma}^2_{i,X}\\ \hat{\varsigma}^2_{i,Y}\\ \hat{\varsigma}_{i,XY}\end{pmatrix} - \begin{pmatrix}\varsigma^2_{i,X}\\ \varsigma^2_{i,Y}\\ \varsigma_{i,XY}\end{pmatrix}\right] \to_d N_3\left[\mathbf{0}_3, \begin{pmatrix} 2\varsigma^4_{i,X} & 2\rho^2\varsigma^2_{i,X}\varsigma^2_{i,Y} & 2\rho\varsigma^3_{i,X}\varsigma_{i,Y} \\ 2\rho^2\varsigma^2_{i,X}\varsigma^2_{i,Y} & 2\varsigma^4_{i,Y} & 2\rho\varsigma_{i,X}\varsigma^3_{i,Y} \\ 2\rho\varsigma^3_{i,X}\varsigma_{i,Y} & 2\rho\varsigma_{i,X}\varsigma^3_{i,Y} & (1+\rho^2)\varsigma^2_{i,X}\varsigma^2_{i,Y} \end{pmatrix}\right].$$

Let $\sigma^2_i = (n_i\phi_i)^{-1} = 2(\varsigma^2_{i,X} + \varsigma^2_{i,Y} - 2\varsigma_{i,XY})/n_i$ and

$$\phi_i = g(\varsigma^2_{i,X}, \varsigma^2_{i,Y}, \varsigma^2_{i,XY}) = [2(\varsigma^2_{i,X} + \varsigma^2_{i,Y} - 2\varsigma^2_{i,XY})]^{-1},$$

with $\hat{\phi} = g(\hat{\varsigma}^2_{i,X}, \hat{\varsigma}^2_{i,Y}, \hat{\varsigma}^2_{i,XY})$. Taking derivatives,

$$\nabla g(\varsigma^2_{i,X}, \varsigma^2_{i,Y}, \varsigma^2_{i,XY}) = -\frac{1}{2}(\varsigma^2_{i,X} + \varsigma^2_{i,Y} - 2\varsigma^2_{i,XY})^{-2}(1, 1, -2)^T$$

so we have, by the delta Method,

$$\sqrt{m_i}(\hat{\phi}_i - \phi) \to_d N\left(0, \frac{2[\varsigma^4_{i,X} - 8\rho\varsigma^3_{i,X}\varsigma_{i,Y} + 2(1+2\rho^2)\varsigma^2_{i,X}\varsigma^2_{i,Y} - 8\rho\varsigma_{i,X}\varsigma^3_{i,Y} + \varsigma^4_{i,Y}]}{4(\varsigma^2_{i,X} + \varsigma^2_{i,Y} - 2\rho\varsigma_{i,X}\varsigma_{i,Y})^4}\right).$$